United States Court of Appeals for the Federal Circuit

MID CONTINENT STEEL & WIRE, INC.,
Plaintiff-Appellee

v.

UNITED STATES,
Defendant-Appellee

PT ENTERPRISE INC., PRO-TEAM COIL NAIL ENTERPRISE INC., UNICATCH INDUSTRIAL CO., LTD., WTA INTERNTIONAL CO., LTD., ZON MON CO., LTD., HOR LIANG INDUSTRIAL CORPORATION, PRESIDENT INDUSTRIAL INC., LIANG CHYUAN INDUSTRIAL CO., LTD.,
Defendants-Appellants

2021-1747

Appeal from the United States Court of International Trade in Nos. 1:15-cv-00213-CRK, 1:15-cv-00220-CRK, Judge Claire R. Kelly.

Decided: April 21, 2022

ADAM H. GORDON, The Bristol Group PLLC, Washington, DC, argued for plaintiff-appellee. Also represented by PING GONG.

MIKKI COTTET, Appellate Staff, Civil Division, United States Department of Justice, Washington, DC, argued for defendant-appellee. Also represented by BRIAN M. BOYNTON, JEANNE DAVIDSON, PATRICIA M. MCCARTHY; VANIA WANG, Office of the Chief Counsel for Trade Enforcement and Compliance, United States Department of Commerce, Washington, DC.

NED H. MARSHAK, Grunfeld, Desiderio, Lebowitz, Silverman & Klestadt LLP, New York, NY, argued for defendants-appellants. Also represented by MAX F. SCHUTZMAN; DHARMENDRA NARAIN CHOUDHARY, ANDREW THOMAS SCHUTZ, Washington, DC.

Before NEWMAN, LOURIE, and TARANTO, *Circuit Judges*.

TARANTO, *Circuit Judge*.

In 2015, the United States Department of Commerce issued an antidumping duty order covering steel nails from Taiwan. In 2019, we ordered a remand to Commerce for further explanation of one aspect of the methodology it had adopted to determine whether there was "a pattern of export prices . . . that differ significantly among purchasers, regions, or periods of time" under 19 U.S.C. § 1677f-1(d)(1)(B)(i). *Mid Continent Steel & Wire, Inc. v. United States*, 940 F.3d 662, 675 (Fed. Cir. 2019) (*CAFC 2019 Op.*). The present appeal involves Commerce's redetermination on remand from our 2019 decision.

In this proceeding, as in others, Commerce, in order to assess the significance of the difference between the prices of two groups of sales, stated that it was using a widely known statistical measure called the Cohen's *d* coefficient. As applied to groups of sales, that coefficient is a ratio whose numerator is the difference between means of the prices of the two groups and whose denominator is a figure, reflecting the general dispersion of the pricing data, that

serves as a benchmark against which to judge the significance of the difference stated in the numerator. Commerce used, for that benchmark, a figure based on the standard deviations of the prices in the two groups; it squared the standard deviations of the prices of each group (yielding the variances), added them together and divided by two, then took the square root. The middle step—adding together and dividing by two—is "simple averaging," which gives equal weight in the average to each group, even if they are very different in size (*e.g.*, if the first group reflects sales of 5 units and the second group reflects sales of 95 units). A "weighted average" approach, in contrast, would, at the middle step, assign weights proportionate to each group's share of the total (*e.g.*, multiplying the first group's variance by 5 and the second by 95, then dividing the sum by 100, thus giving 5/100 weight to the first group and 95/100 weight to the second group). In 2019, we held that Commerce did not adequately explain why it was reasonable to use simple averaging. *Id.* at 673–75. On remand from our decision, Commerce again chose to use simple averaging for its version of a Cohen's *d* denominator.

The Court of International Trade (Trade Court) upheld Commerce's decision. *Mid Continent Steel & Wire, Inc. v. United States*, 495 F. Supp. 3d 1298, 1308 (Ct. Int'l Trade 2021) (*CIT 2021 Op.*). The Taiwanese producers and exporters of the steel nails at issue appeal. We conclude that the relevant statistical literature cited by Commerce uniformly uses weighted averaging in the Cohen's *d* denominator calculation and that Commerce has not offered a reasonable justification for its departure from the cited literature. We therefore vacate the Trade Court's decision and require a remand to Commerce for further consideration of its methodology for applying § 1677f-1(d)(1)(B)(i) here.

I

A

In an antidumping duty investigation, when Commerce seeks to determine whether the foreign-originated merchandise of a foreign producer or exporter is being sold in the United States at less than fair value, *see* 19 U.S.C. § 1673, it must compare the home-country "normal value" (often the sale price in the home country) with the actual or constructed "export price" reflecting the price at which the merchandise is sold into the United States. *CAFC 2019 Op.*, 940 F.3d at 665. That comparison usually calls for use of an "average-to-average" method. When the normal value is based on home-country sales prices of a foreign producer or exporter who is a respondent in the proceeding, the average-to-average method compares "the weighted average of the respondent's sales prices in its home country during the investigation period to the weighted average of the respondent's sales prices in the United States during the same period." *Stupp Corp. v. United States*, 5 F.4th 1341, 1345 (Fed. Cir. 2021); *CAFC 2019 Op.*, 940 F.3d at 666; *see also* 19 U.S.C. § 1677f-1(d)(1); 19 C.F.R. § 351.414(b)(1), (c)(1). But that average-to-average comparison is not the only authorized method: two other methods are authorized, of which one is at issue here.

The statute permits comparisons on a "transaction-to-transaction" basis in unusual circumstances, 19 U.S.C. § 1677f-1(d)(1)(A)(ii); 19 C.F.R. § 351.414(c)(2), but that method is not at issue here. What is at issue is a third method authorized by Congress under certain circumstances—an "average-to-transaction" method. This method calls for the "weighted average of normal values" in the home country to be compared to the "export values (or constructed export values) of individual transactions" in the United States. 19 U.S.C. § 1677f-1(d)(1)(B); 19 C.F.R. § 351.414(b)(3). The object is to uncover "targeted" dumping, a label for an exporter's unduly low pricing in

portions (less than all) of its overall U.S. sales, which would be "masked" (offset) by the exporter's other, higher-priced sales if only overall averages are considered. *See Stupp*, 5 F.4th at 1345. Congress directed that Commerce may use the "average-to-transaction" method only if

> (i) there is a pattern of export prices (or constructed export prices) for comparable merchandise that differ significantly among purchasers, regions, or periods of time, and
>
> (ii) the administering authority explains why such differences cannot be taken into account using a method described in paragraph (1)(A)(i) [average-to-average] or (ii) [transaction-to-transaction].

19 U.S.C. § 1677f-1(d)(1)(B).

The statute does not specify how Commerce should determine whether those conditions are met. *Stupp*, 5 F.4th at 1346. Starting in 2014, Commerce has used a two-stage "differential pricing" analysis. *See* Differential Pricing Analysis; Request for Comments, 79 Fed. Reg. 26,720, 26,722, (May 9, 2014) (*Differential Pricing RFC*); *see also Stupp*, 5 F.4th at 1346–48. The first stage of that process corresponds to the inquiry in paragraph (i)—whether "there is a pattern of export prices . . . that differ significantly among purchasers, regions, or periods of time"—and itself has two parts: the "Cohen's *d* test," followed by the "ratio test." *Differential Pricing RFC* at 26,722–23. The second (final) stage involves a "meaningful difference" assessment to make the determination required in paragraph (ii). *Id.* The present case involves the Cohen's *d* test—the first part of the first stage of Commerce's overall differential pricing analysis.

Under the method as described in 2014, Commerce, considering all sales in the United States by an exporter, is to select a specific purchaser, region, or period of time, form a "test group" consisting of all the exporter's sales meeting

that criterion, and put all the exporter's remaining U.S. sales in the "comparison group." *Id.* at 26,722. That is, Commerce is to compare sales to one purchaser to sales to all others, sales in one region to sales in all others, and sales in one period to sales in all others—in fact, to do so for each purchaser, each region, and each period. For each such test group, Commerce is to compute the Cohen's *d* coefficient by comparing the average price of sales within the test group to the average price of sales within the corresponding comparison group. *Id.*[1] How Commerce did that comparison to calculate the Cohen's *d* in this matter—which appears to be representative of its general approach—is the subject of the dispute before us.

Commerce explained that it started with the following formula from Cohen's textbook to calculate *d*:

$$d = \frac{m_A - m_B}{\sigma}$$

J.A. 1079 (quoting, with font changes, Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* 20 (2d ed. 1988) (*Cohen*)).[2] In that formula, m_A is the mean of the test group (here, the weighted average of the prices of sales in

---

[1] "The Department calculates the Cohen's d coefficient with respect to comparable merchandise if the test and comparison groups of data each have at least two observations, and if the sales quantity for the comparison group accounts for at least five percent of the total sales quantity of the comparable merchandise." *Id.*

[2] It appears that Commerce may have used the "two-tailed" version of the test to account for differences in either direction ($m_A > m_B$ or $m_A < m_B$), taking the absolute value of the coefficient, which is not shown in the formula in the text. *See Stupp*, 5 F.4th at 1346; *Cohen* at 20. That choice is not in dispute here, and the issue before us is unaffected by the presence or absence of absolute value signs in the formula.

the group), m_B is the mean of the comparison group (here, the weighted average of the prices of sales in that group), and σ is "the standard deviation of either population [the test group or the comparison group] (since they are assumed equal)." *Cohen* at 20. Where, as here, the groups consist of sales at known prices, $m_A - m_B$ is in price units (*e.g.*, dollars per kilogram), and so is σ, so the ratio d is a pure (unitless) number.

Commerce then changed the denominator to a figure, also drawn from *Cohen*, designed to be applied when the two groups, though of the same size, have different standard deviations. Specifically, for this new denominator σ′, Commerce used the following formula:

$$\sigma' = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

J.A. 1080 (quoting *Cohen* at 44). In this formula, σ_A^2 and σ_B^2 are the squared standard deviations (variances) of the prices in the test and comparison groups, respectively. The simple average is used under the square-root sign (with no weighting by the sizes of groups A and B), reflecting the fact that, in the situation addressed in the section of *Cohen* containing this formula, groups A and B are of the same size: "$n_A = n_B$." *Cohen* at 43. This formula involves "pooling" the data from the two groups, and the name "pooled standard deviation" is used for both the above formula and also the variation where a weighted average is used instead of a simple average. *E.g.*, *CIT 2021 Op.*, 495 F. Supp. 3d at 1300; *see also CAFC 2019 Op.*, 940 F.3d at 673 (referring to the expression as the "pooled variance" because σ_A^2 and σ_B^2 are the variances of the prices in the two groups).

The disputed feature of the formula is that it does not use the size of the groups to weight the two figures (squared standard deviations, *i.e.*, variances) being averaged. It is undisputed that, when the groups are of the same size, simple averaging equals weighted averaging.

But Commerce used the formula without group-size weighting even when, unlike in the situation described in the *Cohen* section from which the formula is borrowed, the groups are of different sizes. In that circumstance, it is undisputed, simple averaging does not equal weighted averaging. Commerce noted: "To be sure, the use of a simple versus weight[ed] average yields very different results." J.A. 667.

The steps following the calculation of Cohen's *d* in Commerce's analysis are not in dispute. Nor, we note, has Commerce relied on those steps to help justify the simple-averaging choice it has made for the denominator at the first step. We briefly summarize the remaining steps.

Upon calculating *d* for a test group of sales, Commerce described the test group as having "passed" the Cohen's *d* test if *d* for that group exceeded 0.8, *i.e.*, if the difference in means was at least 80% of the pooled standard deviation. *See Mid Continent Steel & Wire, Inc. v. United States*, 219 F. Supp. 3d 1326, 1338–39 (Ct. Int'l Trade 2017) (*CIT 2017 Op.*).[3] Commerce then computed, for the sales of the subject merchandise of a given respondent, the ratio of (a) the total value of those sales which were part of any group that passed the Cohen's *d* test (whether by a purchaser, region, or period comparison) to (b) the total value of all the respondent's sales being studied by Commerce. *Id.* at 1343 n.24. Because that "ratio test" produced a ratio between 33 and 66 percent in this matter, Commerce tentatively decided to use average-to-transaction comparisons in part. *See CAFC 2019 Op.*, 940 F.3d at 671–72.

---

[3] A "pass" thus indicates that the test group's prices are sufficiently different from the comparison group's prices to contribute to a finding of targeted dumping. In this way, the label means the opposite of the word's usual connotation of success in avoiding trouble.

To make its final determination whether to use an average-to-transaction method, Commerce asked, pursuant to § 1677f-1(d)(1)(B)(ii), whether the pricing differences found "cannot be taken into account using" average-to-average or transaction-to-transaction comparisons. For that determination, Commerce asked whether using a comparison other than average-to-transaction would make a "meaningful difference" in the result. Commerce found that there would be such a difference and so adopted the average-to-transaction method. *See CAFC 2019 Op.*, 940 F.3d at 672.

B

1

In response to a petition by Mid Continent Steel & Wire, Inc., Commerce initiated an antidumping duty investigation of certain steel nails from Taiwan and certain other countries. *See CAFC 2019 Op.*, 940 F.3d at 665. The investigation of nails from Taiwan—for the period April 1, 2013, to March 31, 2014—was broken out separately, and Commerce selected PT Enterprises, Inc. and its affiliated producer Pro-Team Coil Nail Enterprise Inc. as mandatory respondents. In May 2015, Commerce issued an affirmative final determination of less-than-fair-value sales in the United States and determined that the appropriate weighted-average dumping margin for those respondents was 2.24%. *Certain Steel Nails from Taiwan: Final Determination of Sales at Less Than Fair Value*, 80 Fed. Reg. 28,959, 28,961 (Dep't of Commerce May 20, 2015) (Final Determination). Following the International Trade Commission's affirmative determination of material injury to a domestic injury, Commerce issued an antidumping duty order. In 2017, following an appeal to the Trade Court, Commerce revised the dumping margin for the respondents to 2.16%. The all-others rate was also set at 2.16%.

Those respondents and other Taiwanese producers and exporters (collectively, PT) and Mid Continent brought

actions in the Trade Court to challenge Commerce's determination. The Trade Court sustained Commerce's application of the Cohen's *d* test in determining whether "there is a pattern of export prices . . . for comparable merchandise that differ significantly among purchasers, regions, or periods of time," 19 U.S.C. § 1677f-1(d)(1)(B)(i), and in particular approved Commerce's decision "to use a simple average to calculate the pooled standard deviation in the Cohen's d test of the differential pricing analysis." *CIT 2017 Op.*, 219 F. Supp. 3d at 1330. In 2019, we mostly affirmed the Trade Court's decision, but we vacated it in part, holding that Commerce's explanation of its use of "a simple average, rather than a weighted average, to calculate the pooled variance used in the Cohen's *d* calculation" was insufficient, requiring a remand to Commerce "for further explanation." *CAFC 2019 Op.*, 940 F.3d at 673, 675.

Specifically, we noted that (1) "Commerce said that it was simply using a widely accepted statistical test; yet it did not acknowledge that the only cited literature source for the relevant aspect of the test itself calls for the use of weighted averages"; (2) Commerce's statement that weighted averaging produces "skewing" was a "mere conclusion" without independent explanation of what the statute calls for; (3) Commerce's rebuttal of PT's argument against the simple average was unsupported and also was not itself an affirmative argument for simple averaging; and (4) Commerce's "predictability" concern seemed tied to the manipulability of reporting sales by number of transactions and Commerce did not indicate why the concern would be present if the average used weighting by quantities or weight of nails sold (nails seemingly being priced per kilogram). *Id.* at 674 (cleaned up). We did not preclude Commerce from making the same decision on remand if it supplied adequate reasoning in support. *Id.* at 675.

2

In December 2019, the Trade Court remanded the matter to Commerce in accordance with our decision. In early March 2020, Commerce issued a draft redetermination decision, again opting to use the simple average to calculate the pooled standard deviation, J.A. 660–76, and attaching portions of three statistics references: *Cohen*, J.A. 723–61; Paul D. Ellis, *The Essential Guide to Effect Sizes* (2010) (*Ellis*), J.A. 678–721; and Robert Coe, *It's the Effect Size Stupid: What Effect Size Is and Why It Is Important*, Paper presented at the Annual Conf. of British Educational Research Ass'n (Sept. 2002) (*Coe*), J.A. 763–73.

In response, PT submitted comments in mid-March 2020, J.A. 780–1004, arguing that "use of simple averaging is both mathematically and statistically inaccurate," J.A. 781. PT pointed to sections of *Cohen* (at 67), of *Coe* (at 6), and of *Ellis* (at 10, 26, 27), all of which set forth formulas that clearly use weighted averages when comparing groups that have both different sizes and different standard deviations (and hence variances). J.A. 790–98.[4] PT proposed a modification, under which the variances of the two groups (test group, comparison group) are weighted by the total weight, in kilograms, of the goods in each group, so the denominator would be

$$\sqrt{\frac{W_a}{W_a + W_b}\sigma_a^2 + \frac{W_b}{W_a + W_b}\sigma_b^2}$$

J.A. 791–92. In that formula, W_a and W_b are the kilogram weights of the test-group goods and comparison-group goods, respectively (and σ_a^2 and σ_b^2 again refer to the variances of the sale prices in the test and comparison groups,

---

[4] The *Coe* reference, at 6 (question 7), is the reference discussed in our 2019 opinion. *CAFC 2019 Op.*, 940 F.3d at 673–74.

respectively). This formula differs in minor ways from the specific formulas in *Cohen*, *Coe*, and *Ellis*, which involve details of weighted averaging appropriate for sampling when not all population data is known. Commerce did not object to PT's formula on the ground that it departed from those models, but rather on the ground that it used weighted averages rather than simple averages.

In May 2020, Mid Continent submitted comments arguing for the simple-average approach. J.A. 1005–70. It included in its comments a discussion of a portion of *Cohen* to which Commerce, in its draft redetermination, had not pointed. J.A. 1022–24 (citing *Cohen* at 360–61). Mid Continent pointed to a statement in *Cohen*—discussing an example involving a researcher's creating equal-size samples of the groups under study, even though some of the groups are a much smaller share of the overall population than the others—about treating a group's characteristic as an "abstract effect quite apart from the relative frequency with which that effect . . . occurs in the population." *Id.*

In June 2020, Commerce published its final redetermination. J.A. 1073–1121. Commerce continued to use a simple average, and it "provid[ed] further explanation of [its] methodology as requested." J.A. 1073. Commerce explained that to determine whether there was a pattern of export prices that "differ significantly" among purchasers, regions, or periods, it used the widely accepted Cohen's *d* test to measure the "effect size" on price associated with sales to certain purchasers, in certain regions, or during certain periods of time, and it relied on *Ellis*, *Cohen*, and *Coe* for elaboration. *See* J.A. 1077–80. It noted that the denominator of the Cohen's *d* coefficient was a "yardstick to gauge the significance of the difference of the means," J.A. 1079, and it stated that the statistical literature presented different methods for computing the denominator, "including the square root of the simple average of the variances within each group," J.A. 1080 (citing *Cohen* at 44).

To justify its decision to use the simple average to calculate the denominator, Commerce wrote:

> [T]he purpose of Commerce's Cohen's *d* test is to determine whether U.S. prices differ significantly among purchasers, regions, or time periods – *i.e.*, do prices to each purchaser, region, or time period differ significantly from all other prices of the comparable merchandise. Although these are all prices in the U.S. market made by the respondent, this analysis requires that these prices be subdivided into separate distinct groups to consider separately whether the respondent's pricing behavior for sales to one specific group differs from its pricing behavior for all other sales. In other words, these prices, all of which are used to evaluate: 1) a respondent's pricing behavior in the U.S. market; and 2) whether the respondent is dumping, are now considered to represent two distinct pricing behaviors which may differ significantly. For the purpose of this particular analysis, Commerce finds that these two distinct pricing behaviors are separate and equally rational, and each is manifested in the individual prices within each group. Therefore, each warrants an equal weighting when determining the "standard deviation" used to gauge the significance of the difference in the means of the prices of comparable merchandise between these two groups. Because Commerce finds that each of these pricing behaviors are equally genuine when considering the distinct pricing behaviors between a given purchaser, region, or time period and all other sales, an equal weighting is justified when calculating the "standard deviation" of the Cohen's *d* coefficient. To do otherwise and use an average weighted by sales volume, sales value, or number of transactions would give preference to one pricing behavior over the other, and therefore would bias

> the "yardstick" by which Commerce measures the observed difference in prices between the test and comparison groups.

J.A. 1080–81.

In responding to comments, Commerce referred to the "abstract effect" idea invoked by Mid Continent. J.A. 1112, 1116–17. It also pointed to the difference between this context, in which Commerce has the complete population data pool (and each pairwise comparison involves the entire population), and the context of the cited literature involving sampling from a population. J.A. 1109. Commerce further said that PT's challenge of the simple average relied on conclusory allegations of "skewed" results, J.A. 1081, incorrect assumptions about the relationship between standard deviation and group size, J.A. 1083–84, and "cherry-picked" data, J.A. 1084–85. It added that the simple average provides "predictability" because "the importance given to each pricing behavior will be the same for all products," and it concluded that the use of a simple average was "not only a reasonable approach but a more accurate and consistent measurement." J.A. 1087.

3

The matter returned to the Trade Court. PT submitted comments that included extensive attachments containing the sales information before Commerce and figures that, according to PT, showed why weighted averaging is substantially better than simple averaging at capturing those instances in which a test group's prices are noticeably outside the dispersion of prices generally. J.A. 1122–1373. The government responded, arguing, among other things, that PT failed to exhaust administrative remedies as to some of what PT presented. J.A. 1397–1428.

In January 2021, the Trade Court sustained Commerce's determination. *CIT 2021 Op.*, 495 F. Supp. 3d at 1300. It accepted Commerce's explanation that a weighted

average would "inappropriately move the pooled standard deviation toward the pricing behavior of either the test or comparison group," *id.* at 1304, and agreed that an equal weighting was justified because the prices in each test and comparison group "separately and equally represent the respondent's pricing behavior," *id.* at 1308 (quoting J.A. 1108). The Trade Court did not refer to the "abstract effect" idea invoked by Mid Continent and Commerce.[5]

PT timely appealed to this court. We have jurisdiction under 28 U.S.C. § 1295(a)(5).

II

A

We review Commerce's decisions using the same standard of review applied by the Trade Court, while carefully considering the Trade Court's analysis. *CAFC 2019 Op.*, 940 F.3d at 667. Commerce's selection of a methodology for implementing the statutory directive of § 1677f-1(d)(1)(B) is "an interpretation of that statutory language" that we review for reasonableness. *Stupp*, 5 F.4th at 1352–53; *see Ningbo Dafa Chem. Fiber Co. v. United States*, 580 F.3d 1247, 1256 (Fed. Cir. 2009) ("It is well established that statutory interpretations articulated by Commerce during its antidumping proceedings are entitled to judicial deference under *Chevron*." (cleaned up)).

---

[5] The Trade Court reached its conclusion without having to determine which if any submissions by PT were objectionable under the exhaustion requirement, because the court concluded that all of the submissions were, in any event, answered by the just-noted rationale. *Id.* at 1306–08. Our decision does not rely on the materials that were the subject of the exhaustion dispute, which we therefore need not address.

"Commerce has discretion to make reasonable choices within statutory constraints." *CAFC 2019 Op.*, 940 F.3d at 667; *see also Stupp*, 5 F.4th at 1353, 1354. Commerce's "special expertise in administering antidumping duty law" is one recognized basis for the "significant deference" embodied in the reasonableness standard. *Ningbo Dafa*, 580 F.3d at 1256; *see also Wheatland Tube Co. v. United States*, 495 F.3d 1355, 1361 (Fed. Cir. 2007). Expertise enables an agency to identify a reasonable interpretation and to set forth an adequate justification for choosing it over others, but it remains a judicial obligation to ensure that the agency has done so, while avoiding judicial usurpation of agency authority to make pertinent factual and policy determinations. *See Burlington Truck Lines, Inc. v. United States*, 371 U.S. 156, 167–69 (1962); *CS Wind Vietnam Co. v. United States*, 832 F.3d 1367, 1377 (Fed. Cir. 2016). For us to fulfill that obligation, we must ensure that Commerce provides "an explanation that is adequate to enable the court to determine whether the choices are in fact reasonable, including as to calculation methodologies." *CAFC 2019 Op.*, 940 F.3d at 667; *Stupp*, 5 F.4th at 1357.

Last year, in *Stupp*, we held that Commerce had provided an inadequate explanation of the reasonableness of its use of Cohen's *d* in its differential-pricing analysis in circumstances where that use seemingly departed from what the statistical literature taught. *Stupp*, 5 F.4th at 1357–60. What was unjustified there was Commerce's use of Cohen's *d* "in adjudications in which the data groups being compared are small, are not normally distributed, and have disparate variances." *Id.* at 1357. We remanded for further consideration.

On the record presented to us here, we do the same, focusing on a different feature of Commerce's use of Cohen's *d*. We hold that Commerce has not adequately justified its adoption of simple averaging for the Cohen's *d* denominator. Commerce has departed from the methodology described in all the cited statistical literature

governing Cohen's *d*, but it has not justified that departure as reasonable. We again remand for further consideration.

B

1

Commerce recognized that the function of the denominator in the Cohen's *d* coefficient is to be a "yardstick to gauge the significance of the difference of the means" of the sales prices of the test and comparison groups. J.A. 1079. The numerator of Cohen's *d* is the difference in weighted average sales prices between the test and comparison groups. Without further context, *i.e.*, without a basis for comparison, it is impossible to say whether that difference is "significant," under 19 U.S.C. § 1677f-1(d)(1)(B)(i) or otherwise. The central purpose of using the Cohen's *d* ratio is to provide the missing basis of comparison—the "yardstick." Cohen's *d* relates, by division, the difference in mean prices of the two particular groups to a figure representing the magnitude of differences in (dispersion of) the prices in the data pool more generally. *See CAFC 2019 Op.*, 940 F.3d at 671. If the mean-price difference is large enough compared to the more general dispersion measure (*i.e.*, the ratio of the two is at least 0.8), "Commerce deems the sales prices in the test group to be significantly different from the sales prices in the comparison group." *Stupp*, 5 F.4th at 1347; *see Differential Pricing RFC* at 26,722 ("The Department finds that the difference is significant, and that the sales of the test group pass the Cohen's d test, if the calculated Cohen's d coefficient is equal to or exceeds the large threshold.").

The cited literature makes clear that one way to form the more general data-pool dispersion figure for the denominator—seemingly the preferred way if the full set of population data is available—is to use the standard deviation for the entire population. But the references recognize that entire population data may be unavailable, in which case an alternative is needed, and the alternative is chosen with

the object of estimating (approximating) the unavailable population standard deviation. Thus, *Ellis* states:

> To calculate the difference between two groups we subtract the mean of one group from the other ($M_1 - M_2$) and divide the result by the standard deviation (SD) *of the population from which the groups were sampled*. The only tricky part in this calculation is figuring out the population standard deviation. If this number is unknown, some approximate value must be used instead.

Ellis at 10 (emphasis added). *Coe* presents the formula for measuring effect size as

$$\frac{[Mean\ of\ experimental\ group] - [Mean\ of\ control\ group]}{Standard\ Deviation}$$

and then states:

> The "standard deviation" is a measure of the spread of a set of values. Here it refers to the standard deviation *of the population from which the different treatment groups were taken*. In practice, however, this is almost never known, so it must be estimated either from the standard deviation of the control group, or from a "pooled" value from both groups . . . .

Coe at 2 (emphasis added). And *Cohen* similarly indicates that the ideal denominator is the full population's standard deviation, which may be approximated by a pooled estimate. *See Cohen* at 27 (dividing by "the common within-population standard deviation"); *Cohen* at 67 (noting that the denominator is "the usual pooled within sample estimate of the population standard deviation"—indicating that the pooling method, based on the standard deviations of each of the two groups, aims to estimate the standard deviation of the overall population). When the full population data set is unavailable, all of the cited literature points to use of a "pooled standard deviation" of the two particular

groups at issue to form the denominator. *Cohen* at 67; *Ellis* at 10, 26–27; *Coe* at 6.

In this matter, Commerce did not use the standard deviation of all the data for its denominator. It made that choice even while recognizing that it had the full set of data for U.S. sales for the period Commerce was reviewing. J.A. 1109 ("Commerce's analysis is based on all of the U.S. sales data for the respondent . . . . Commerce does not sample the respondent's U.S. sales data used in the Cohen's *d* test, and the calculated means and variances of the U.S. prices are the actual values of the entire population of U.S. sales and are not estimates of those values."). Indeed, in each test-group/comparison-group pair, the test and comparison groups together make up "the entire universe, *i.e.*, population, of the available data," J.A. 1115, because for each test group, the comparison group is all other sales data.

Rather than use the population standard deviation in the denominator, Commerce used a "pooled standard deviation," pooling the standard deviations for each pair of test and comparison groups. As discussed above, it used simple averaging to do the pooling—even where the test and comparison groups have different sizes. In making that choice to use simple averaging, however, Commerce departed from, rather than followed, the cited statistical literature. As we have described above, Commerce's formula for the denominator,

$$\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

comes from a section of *Cohen* that addresses a situation in which the two groups at issue are of the same size. *Cohen* at 43–44; *id.* at 43 ("CASE 2: $\sigma_A \neq \sigma_B, n_A = n_B$"). By contrast, when the sampled groups have unequal sizes, the cited literature uniformly teaches use of a pooled standard deviation estimate that involves weighted averaging. *See Cohen* at 67; *Ellis* at 26–27; *Coe* at 6.

The section of *Cohen* (at 359–61) cited by Mid Continent and Commerce for its "abstract effect" language is no exception. It nowhere recites use of a simple average for calculating a pooled standard deviation from groups of unequal size. The discussion in that section involves *f*, an effect size index that is related to, but not the same as, the Cohen's *d* coefficient, applicable when there are arbitrarily many groups to compare, rather than just two. *See Cohen* at 274–80. It expressly sets forth a simple average formula for when the groups are equal in size but a weighted average formula for when the groups are of different size. *Id.* at 359–60. The language of "abstract effect" is used in a discussion of forming certain equal-size groups for the comparative analysis: in the example given, if the object was to identify differences in viewpoint on a topic (attitudes toward the United Nations) among three groups (Jews, Protestants, Catholics), the researcher could form equal groups even though random sampling from a population would produce different-size groups. *Id.* at 360–61. Nothing in the section applies simple averaging to pooled standard deviation estimates for different-size groups.

2

Commerce offered one principal reason for departing from the teaching of all the cited statistical literature. It said that the data in each group (the test and comparison groups) represent "equally rational" and "equally genuine" pricing choices and that, therefore, each group "warrants an equal weighting" for calculating the pooled standard deviation. J.A. 1080–81. We see no basis for questioning, here or generally, the premise of equal rationality of the pricing behavior (and equal genuineness, if that is different, which is not clear). But Commerce has not offered an adequate explanation of why that premise supports the particular step Commerce must justify: a choice of how to form the denominator in the Cohen's *d* formula.

The fact that the seller is acting rationally and genuinely in its pricing choices in both the test and comparison groups provides no apparent reason for assigning equal weight to each group's standard deviation when computing the pooled standard deviation. The rationality and genuineness of the seller's pricing choices have no evident connection to the undisputed purpose of the denominator figure—to provide a dispersion figure for the more general pool that serves as a yardstick for deciding on the significance of the difference in mean prices of the two groups. Both the numerator and denominator take the behavior as a given and form certain statistical measures from the objective data that are then related in the ratio that is Cohen's *d*. Commerce has not identified anything in the statistical measure at issue that depends on considerations of rationality and genuineness of the conduct that gave rise to the objective data. Indeed, Commerce has not shown that the numerous real-world examples used in *Cohen* to illustrate the methods taught are different in the respect Commerce now features, *i.e.*, Commerce has not shown that the *Cohen* examples (generally or, perhaps, ever) involve sampled groups of data that reflect behavior that is *not* "rational" and "genuine." Thus, Commerce has not adequately justified, through its central rationale, its departure from the statistical literature's description of the Cohen's *d* coefficient.

Commerce also asserted that a simple average provides "predictab[ility]" in that "the importance given to each pricing behavior will be the same for all products." J.A. 1087. But Commerce did not suggest that this basis would suffice for its denominator choice without the principal basis we have just discussed and found inadequate. And in any event, Commerce has not provided a reasonable explanation for this predictability assertion. It is not clear from Commerce's language, including its "importance given to each pricing behavior" language, what meaning Commerce was ascribing to "predictability" independent of its equality

(of rationality and genuineness) basis. If Commerce was referring, as "predictability" would suggest, to the ability to predict the consequences for the dumping analysis based on the ability to predict the weighting of a sale (the "importance" component of the analysis), Commerce did not explain why simple averaging has greater predictability than weighted averaging (let alone than using the full population's standard deviation for every *d* calculation). The mathematical formulas have no identified elements of discretion, or other components, that distinguish them with respect to prediction. Specifically, Commerce provided no basis for an assertion of lesser "predictability" if weighted averaging is done on the basis of weight (or dollars or units), not transactions, as we discussed in our 2019 opinion. *See CAFC 2019 Op.* at 674. Not having provided an adequate explanation of "predictability," Commerce also did not provide an adequate explanation of what significance this consideration should have in the overall choice of denominator for Cohen's *d*.

In its final redetermination, Commerce invoked the "abstract effect" idea mentioned in the section of *Cohen* discussed above. J.A. 1112, 1116–17. As we have noted, that section does not call for simple averaging for unequal size groups in the denominator of Cohen's *d* or in the formula for the related *f* figure. And Commerce has not explained how such simple averaging could be derived from the "abstract effect" idea itself. We do not understand Commerce, in invoking this idea, to be saying anything other than that the statutory "differ significantly" analysis focuses on the difference between the test and comparison groups for its own sake, rather than for what it indicates about the overall population. One difficulty with this observation is that Commerce has not explained how it affects comparisons, such as those Commerce makes in its differential-pricing analysis, where the groups together make up the entire population (which was not the case in the section of *Cohen* at issue). More broadly and fundamentally, Commerce has

not explained why the fact that the focus is being placed on the difference between the groups distinguishes the teaching of the cited literature—which, as discussed, uses the Cohen's *d* coefficient precisely to provide a yardstick for determining the significance of the difference in group means. Thus, Commerce has not explained why that focus calls for a simple-averaging yardstick figure for determining the significance of the difference when calculating Cohen's *d* (or, even, the *f* statistical measure) for different-size groups.

Commerce observes that the cited literature discusses "sampling" from a population, whereas Commerce has the entire population data and each of its test-comparison group pairs involves the entire population. J.A. 1109. In *Stupp*, we stated that Commerce had not explained how this difference bears on the reasonableness of Commerce's use of Cohen's *d* in certain respects not at issue in the present matter. 5 F.4th at 1360. Here, too, although it is undisputed that sampling for estimation of an unknown overall population figure requires certain minor alterations of the formula for weighted averaging not needed in the present context, *compare*, *e.g.*, *Cohen* at 67, *with* J.A. 792 (PT proposal), Commerce has not explained why the basic choice of weighted averaging of unequal-size groups fails to apply to the present context. The cited literature nowhere suggests simple averaging for unequal-size groups. Indeed, when the entire population is known, the cited literature points toward using the standard deviation of the entire population as the denominator in Cohen's *d*—which Commerce has not done.

3

Commerce's job is not to follow a statistical test as explained in published literature for its own sake, but to implement the statutory mandate to determine when prices of certain groups "differ significantly." 19 U.S.C. § 1677f-1(d)(1)(B)(i). In implementing a statutory mandate, an

agency is not duty-bound to follow published literature when, *e.g.*, the literature is inapplicable to the specific problem before the agency or is not itself well grounded. But here Commerce embraced the Cohen's *d* statistics measure and relied on the literature for that measure in making its statutory significance assessment—and that embrace extends beyond the first step and is the foundation of the remaining steps. After the calculation of Cohen's *d*, the next step in Commerce's analysis is to declare what number is high enough to be significant (constituting "passing" the Cohen's *d* test), and the number it uses is 0.8, the threshold for a "large" effect size stated in *Cohen*. *See Cohen* at 26; J.A. 1080; *Differential Pricing RFC* at 26,722; *Stupp*, 5 F.4th at 1347. The "passing" sales then determine the results of the next "ratio test" step.

In this situation, Commerce needs a reasonable justification for departing from what the acknowledged literature teaches about Cohen's *d*. It has departed from those teachings about how to calculate the denominator of Cohen's *d*, specifically in deciding to use simple averaging when the groups differ in size. And its explanations for doing so fail to meet the reasonableness threshold (a deferential one, in recognition of expertise) for the reasons we have set forth.

We must remand for further proceedings before Commerce in light of the identified deficiencies—as we did in this matter in 2019 regarding the simple-averaging choice and as we did in *Stupp* regarding other aspects of Commerce's use of Cohen's *d*. Commerce must either provide an adequate explanation for its choice of simple averaging or make a different choice, such as use of weighted averaging or use of the standard deviation for the entire population.[6]

---

[6] Mid Continent argues that, if weighted averaging is to be done, the weighting should be based on the number

III

For the foregoing reasons, we vacate the decision of the Trade Court and remand for further proceedings consistent with this opinion.

No costs.

VACATED AND REMANDED

---

of transactions, rather than on a measure of how much is sold (*e.g.*, number of nails, weight of nails, dollars paid). Mid Continent Br. 28–29. But Commerce rejected weighted averaging altogether, so we do not have before us for review a choice of one basis of weighting rather than another. We make two observations relevant to Commerce's consideration of that choice if it adopts weighted averaging on remand. First, when it uses the average-to-average method, Commerce computes average prices by quantity sold, not by transaction. *See* J.A. 1111. Second, in our earlier opinion, we recognized that Commerce had criticized weighting by the number of transactions as susceptible to manipulation, and we noted that weighting by quantity appears to address that issue. *CAFC 2019 Op.*, 940 F.3d at 674.